Learning Ordinal Relationships for Mid-Level Vision

Daniel Zoran
CSAIL, MIT
danielz@mit.edu

Phillip Isola
CSAIL, MIT
phillipi@mit.edu

Dilip Krishnan
Google
dilipkay@google.com

William T. Freeman
Google and CSAIL, MIT
wfreeman@google.com

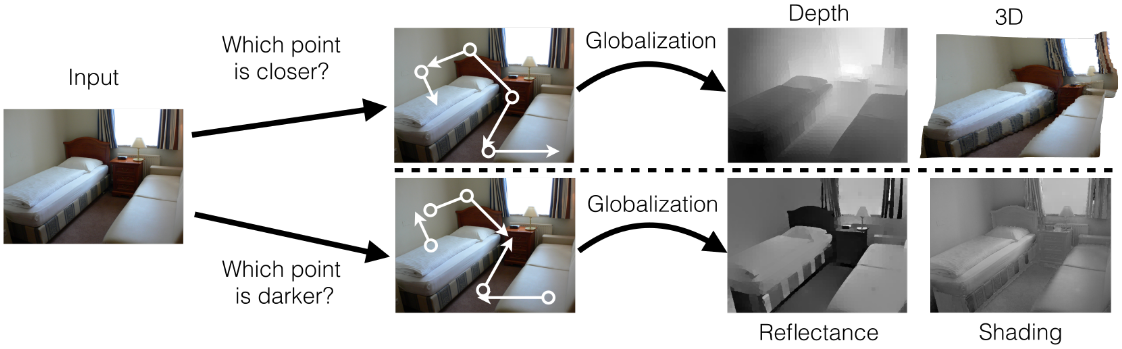

Figure 1: Framework overview: given an input image we choose points on which to make ordinal estimates (*e.g*. Which of two points is closer to the camera? Which has darker surface color?). We train models to perform these estimations. We then globalize these estimates to produce metric estimates of reflectance, shading and depth. Our approach has multiple benefits over direct metric estimation.

Abstract

We propose a framework that infers mid-level visual properties of an image by learning about ordinal relationships. Instead of estimating metric quantities directly, the system proposes pairwise relationship estimates for points in the input image. These sparse probabilistic ordinal measurements are globalized to create a dense output map of continuous metric measurements. Estimating order relationships between pairs of points has several advantages over metric estimation: it solves a simpler problem than metric regression; humans are better at relative judgements, so data collection is easier; ordinal relationships are invariant to monotonic transformations of the data, thereby increasing the robustness of the system and providing qualitatively different information. We demonstrate that this framework works well on two important mid-level vision tasks: intrinsic image decomposition and depth from an RGB image. We train two systems with the same architecture on data from these two modalities. We provide an analysis of the resulting models, showing that they learn a number of simple rules to make ordinal decisions. We apply our algorithm to depth estimation, with good results, and intrinsic image decomposition, with state-of-the-art results.

1. Introduction

Mid-level vision involves the estimation of physical properties of pixels within an image. Examples of such problems are depth extraction, surface normal orientation estimation, intrinsic image decomposition, segmentation and shadow detection/removal [23, 4, 35, 13, 20, 22, 34]. Usually, metric (continuous valued) estimates are required at each pixel location. However, there are many cases where such a quantity may be hard to extract, or even if estimated, may be irrelevant to the task at hand. Consider looking at a piece of paper on top of a table: it would be hard to say how much closer the paper is to the observer, since the metric difference is tiny. An *ordinal* relation between the paper and the table, however, is clearly present. A similar case happens in outdoor scenes: looking at a range of mountains, it is usually easy to say which mountain is in front of which, but metric distances between them may be hard to estimate by sight alone. Another example is the estimation of surface albedo (reflectance) within an image. It can be difficult to give a precise albedo estimate, and this has been shown by experiments on lightness perception [9]. On the other hand, it is natural for an observer to say which of two points has darker or lighter reflectance. Only in rare cases do humans fail at this task, and such cases can be newsworthy, as in the Internet controversy over the color of a particular dress [1].

Ordinal relationships and rankings have been the subject of much research in computer vision [6, 26, 27], machine learning [7] and psychophysics [9]. There is strong empirical evidence that people are better at estimating ordinal relations between points in scenes than estimating metric quantities [33]. Such relationships have several attractive properties that make them suitable for vision applications. They have a simple, discrete three way classification structure (corresponding to the "equality", "less than", and "greater than" relationships). They are invariant to monotonic transformations of the data, making them more robust than metric estimates to variations of illumination, viewpoint or pose. Since humans find them easier to estimate, they can help us collect ground truth data for cases where metric estimates would be hard to acquire.

In this paper we propose a framework that tackles mid-level vision problems by learning from ordinal relationships. We train a deep neural net to estimate the ordinal relationships between point pairs in a given image. These estimates are then aggregated to provide a full explanation of the image (see Figure 1). We demonstrate the system on two modalities: reflectance and depth. We show that the same system architecture can learn from these different modalities and provide an analysis for the kind of decision rules the system may have learnt. We show that we achieve competitive results in depth from single image and state-of-the-art results on intrinsic image decomposition.

2. Related Work

A characterization of mid-level vision problems is due to Marr [23], where localized primitives such as pixel values, edges and corners are used to provide more global descriptions of a scene. The estimation of depth, reflectance and shading, and segmentation are important mid-level vision tasks. We refer to the reader to [38] and [31] for good reviews.

Most work that extracts reflectance and shading from a single image relies on engineered priors [14, 12, 28, 21]. These models usually use cues such as gradient characterizations of reflectance and shading [18] or sparsity of albedo values in a given image [28]. Smoothness priors are also used in most reflectance models. Depth extraction from a single image has proven to be a harder problem, and almost all existing models make strong assumptions such as known illumination or shading [15, 35]. A notable exception is the recent work of Barron and Malik [4], which uses a number of physics-based models to recover reflectance, depth and illumination from a single RGB or RGB-D image. However, the restriction in their case is either the requirement of an external depth map from a depth sensor, or the presence of only a single object in the image (with a mask outlining the silhouette).

Learning-based approaches provide an alternate mechanism to hand-designed priors. However, reliable models for mid-level problems could not be learnt in the past due to the absence of large-scale datasets, except for specific synthetic datasets such as [10]. In the case of depth, the availability of cheap depth sensors has given rise to datasets such as NYU Depth [25]. These open up the possibility of learning priors directly from the data, and recent works give excellent results for depth estimation from a single RGB image [8, 20, 34, 22], showing the power of learnt models. There have been relatively few works on learning intrinsic image decompositions [12, 32] - perhaps due to the lack of data. However, the recently released Intrinsic Images in the Wild (IIW) [5] provides a large amount of data, and several works concurrent with our own have learned from this data [24, 40].

Almost all existing models try to predict metric depth or reflectance values given an input image. An alternative, considered in this paper, is to predict ordinal relationships between pairs of points and use those to recover a metric estimate. Ordinal features have been successfully used in many applications in previous work, such as image correspondence [6], attribute prediction [27], face recognition [29], and texture classification [26]. The concurrent works of [24, 40] also learn ordinal classifiers for reflectance, but do not apply their frameworks to depth.

3. Framework

Our framework has three main components. The first one selects point pairs (i, j) from the image. The second component performs ordinal relationship estimation for each pair: extracting the relevant information from the image and providing a three-way classification. Finally, the third component takes the partial order relationship estimates between points in the image and propagates them to provide a dense depth or reflectance map for the whole image, utilizing priors specific to the task at hand.

We shall now describe the components of our framework. We use the same architecture for both the tasks demonstrated here (intrinsic image decomposition and depth extraction), demonstrating its generality. We note, however, that the framework is modular and individual components can be independently adapted to the task at hand.

3.1. From input image to point pairs

We wish to choose N points from the image and an edge structure $E_{i,j}$ denoting which pairs of points i, j are to be compared. Depth or reflectance discontinuities usually induce changes in the measured intensity within the image, creating strong gradients. The points we want to compare need to be far away from such gradients so as to be centered within relatively homogenous regions. The lines connecting the points, however, should ideally cross such gradients, allowing us to make the decision about the possible cause of the observed change.

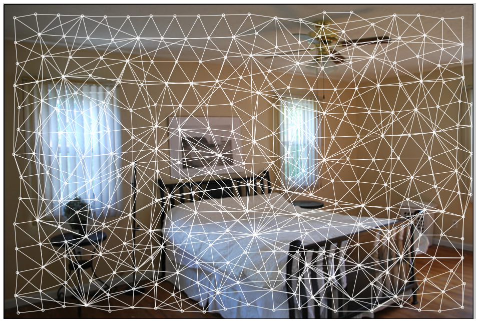

Figure 2: The points and edge structure extracted from a given image: a comparison is estimated on each edge. Note (a) that points rarely straddle strong edges and (b) the long range connectivity. (For this image we used a larger separation between the points for display purposes).

A simple way of finding points that satisfy the above requirements is to use the centers of superpixels [2]. Superpixels naturally follow gradients within images, so their centers tend to be locally far away from strong intensity changes. In order to choose edge structure connecting the points, we find for each point its adjacent neighbors in the superpixel segmentation. The first order neighbors often lie on opposing sides of strong gradients in the image. To allow for longer range interactions, we introduce connections between higher-order neighbors: we extract superpixels at several sizes, find their first order neighbors and snap this edge structure to the fine superpixel grid. See Figure 2 for an example of the resulting points and edge graph. Note that the superpixels are used only for their centers, and no segmentation information is utilized at this stage.

3.2. From point pairs to ordinal estimates

For each edge $E_{i,j}$ connecting points i and j in the image, we need to make a decision about the ordinal relationship between i and j. This can be posed as a three way classification problem where the classes correspond to equality between the points, point i being larger and point j being larger. Here, the word "larger" has a natural interpretation according to the task: it means "darker" for intrinsic image decomposition, and "further away" for depth estimation, for example. Along with this estimate we would want an associated confidence value (namely, the normalized class probabilities).

We train a deep neural network to perform this classification. The network input is comprised of two different contexts: the first is the local appearance of the two points and their local surroundings, the second is a global context, which provides the network the overall image structure and the region of interest (ROI) in which the points reside. Specifically, for local context, we extract a patch around each point, their joint bounding box and provide the network with a mask denoting the points' relative location within the bounding box. For global context, we input a downscaled version of the image, and a mask denoting the ROI: the location of the local context bounding box within the image. See Figure 3 for a depiction of the network input. The network architecture and training procedure are detailed in Section 4.

3.3. From ordinal to metric

The output of the deep network are 3 ordinal relationship probabilities for each pair of considered points. We need to propagate these estimates to all other points in the image, and move from ordinal estimates to metric. To achieve this, we must reconcile the ordering estimates for the points, some of which may be contradictory or ambiguous. Second, we need to find values for points which were not estimated, these are in fact the majority of pixels in the image. For the latter, we use superpixel segmentation and assume that the value throughout each superpixel is constant (yielding a piecewise constant solution in the resulting image). More sophisticated priors such as piecewise planar assumptions may be used [36], but we have found that as long as the superpixels are of small enough size, the constancy assumption gives good performance while speeding up inference.

Given this piecewise constancy assumption, we seek to find a global solution for the values of the points (denoted by x_i), while handling ambiguities in the orderings, incorporating the confidence of the classifier and respecting image information and problem scale. We pose this as a constrained quadratic optimization problem, which we explain

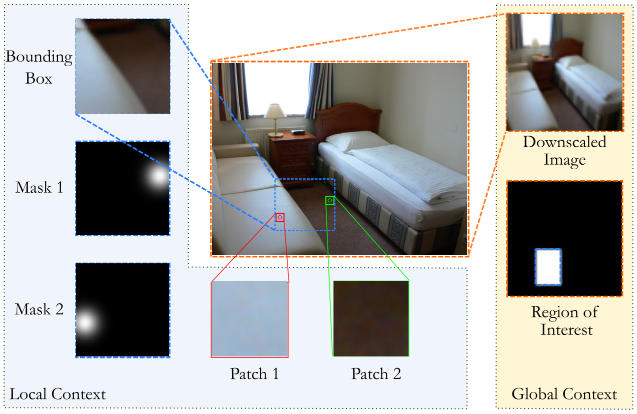

Figure 3: Inputs to the network: we extract a patch around each point of interest (red and green squares on the image). These patches, together with the bounding box (blue) of the patches and masks denoting the relative position of the points within the bounding box, form the "local context" for the network. The "global context" is provided by a downscaled version of the image, along with a Region of Interest (ROI) mask denoting the location of the bounding box within the image.

below. We denote the equality decision using the superscript or subscript eq, the first inequality decision (point i being larger than point j) by gt and the opposite inequality decision by lt.

Our first term handles equality predictions. Since the network predictions are uncertain, we do not want to use hard constraints. Instead, for *each* edge $E_{i,j}$ we wish to minimize the l_2 distance between the estimates of the corresponding points, weighted by the confidence of the equality prediction w^{eq}_{ij}. We wish to minimize:

$$\mathcal{L}_{eq}(\mathbf{x},\mathbf{R}^{eq}) = \sum_{ij} w^{eq}_{ij}(x_i - x_j - R^{eq}_{ij})^2 \tag{1}$$

w.r.t. $\mathbf{x}$ and $\mathbf{R}^{eq}$, where $R^{eq}_{ij} \sim \mathcal{N}(0,\sigma^2_{eq})$ is a scalar slack variable for the ij-th edge. The variance σ^2_{eq} of the normal distribution for R^{eq} is computed from the statistics of the training set. The above can be written in matrix form as follows. Let $\mathbf{A}_{eq} \in \mathbb{R}^{|E|\times N+|E|}$, where $|E|$ is the number of edges, and N the number of points. Each row of $\mathbf{A}_{eq}$ has the entries 1,-1 and -1 in the i, j and $N+p$ columns, respectively (p is the index of ij pair). The weights w^{eq}_{ij} can be represented by a diagonal matrix $\mathbf{W}_{eq} \in \mathbb{R}^{|E|\times|E|}$ in which the (diagonal) entries are w^{eq}_{ij}. Using this, Eq. 1 can be written in matrix form as:

$$\mathcal{L}_{eq}(\mathbf{x},\mathbf{R}^{eq}) = [\mathbf{x}\ \ \mathbf{R}^{eq}]^T\,\mathbf{A}^T_{eq}\mathbf{W}_{eq}\mathbf{A}_{eq}\begin{bmatrix}\mathbf{x}\\ \mathbf{R}^{eq}\end{bmatrix} \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{R}^{eq} \in \mathbb{R}^{|E|}$ are the variables x_i and R^{eq}_{ij} respectively, stacked in vector form. For clarity, here and subsequently $[\mathbf{x}\ \ \mathbf{R}^{eq}]$ is a row vector.

Inequalities are handled in a similar manner. First, let us consider the inequalities corresponding to the prediction that point i is larger than point j. For each edge, we require the estimated difference to be some non negative number R^{gt}_{ij} such that:

$$\mathcal{L}_{gt}(\mathbf{x},\mathbf{R}^{gt}) = \sum_{ij} w^{gt}_{ij}\left(x_i - x_j - R^{gt}_{ij}\right)^2 \tag{3}$$

where w^{gt}_{ij} is the associated confidence for this inequality ($x_i > x_j$), and $R^{gt}_{ij} \sim \mathcal{N}(\mu_{gt},\sigma^2_{gt})$ are the positive-valued slack variables. Note that μ_{gt} is a positive value and that all R^{gt}_{ij} are constrained to be positive. Analogous to the equality constraints Eq. 2, this can be written in matrix form, with a matrix $\mathbf{A}_{gt}$ with entries in each row of value 1,-1 and -1 and a diagonal weight matrix $\mathbf{W}_{gt}$:

$$\mathcal{L}_{gt}(\mathbf{x},\mathbf{R}^{gt}) = \left[\mathbf{x}\ \ \mathbf{R}^{gt}\right]^T\,\mathbf{A}^T_{gt}\mathbf{W}_{gt}\mathbf{A}_{gt}\begin{bmatrix}\mathbf{x}\\ \mathbf{R}^{gt}\end{bmatrix} \tag{4}$$

Finally, we have a third term, $\mathcal{L}_{lt}(\mathbf{x},\mathbf{R}^{lt})$, for the inequality $x_i < x_j$, with matrix $\mathbf{A}_{lt} = -\mathbf{A}_{gt}$, weights $\mathbf{W}_{lt}$ and slack variables $\mathbf{R}^{lt}$.

In addition to the above terms, we also take into account the image structure. Specifically, we want to enforce smoothness with respect to the image. This is achieved with an ℓ_2 smoothness term between adjacent superpixels m and n, weighted by local image gradients (calculated over the mean superpixel luminance L) such that $w^s_{m,n} = \exp(-\frac{1}{\rho}\|L_m - L_n\|^2)$ where ρ controls the sensitivity of the smoothness weights. For non-adjacent superpixels, $w^s_{m,n} = 0$. We use these weights in one of two ways. First, we either directly smooth the values $\mathbf{x}$, giving rise to a smoothness term is of the form:

$$\mathcal{L}_s(\mathbf{x}) = \sum_{ij} w^s_{i,j}(x_i - x_j)^2 + \sum_i b_i x_i \tag{5}$$

where the terms b_i allow more flexibility in the choice of smoothing (see Section 4 for an example). This can again be re-written in matrix form:

$$\mathcal{L}_s(\mathbf{x}) = \mathbf{x}^T\mathbf{A}^T_s\mathbf{W}_s\mathbf{A}_s\mathbf{x} + \mathbf{x}^T\mathbf{b}_s. \tag{6}$$

Combining all the above expressions, the resulting constrained quadratic problem is given by:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{R}^{eq},\mathbf{R}^{gt},\mathbf{R}^{lt}}\ & \lambda_{eq}\mathcal{L}_{eq}(\mathbf{x},\mathbf{R}^{eq}) + \lambda_{gt}\mathcal{L}_{gt}(\mathbf{x},\mathbf{R}^{gt}) + \\ & \lambda_{lt}\mathcal{L}_{lt}(\mathbf{x},\mathbf{R}^{lt}) + \lambda_s\mathcal{L}_s(\mathbf{x}) + \\ \sum_{ij}\ & \left(\frac{(R^{eq}_{ij})^2}{\sigma^2_{eq}} + \frac{(R^{gt}_{ij}-\mu_{gt})^2}{\sigma^2_{gt}} + \frac{(R^{lt}_{ij}-\mu_{lt})^2}{\sigma^2_{lt}}\right) \\ \text{s.t}\ \ & \mathbf{x} > \mathrm{L}, \mathbf{x} < \mathrm{U}, \mathbf{R}^{eq} > 0, \mathbf{R}^{gt} > 0, \mathbf{R}^{lt} > 0\end{aligned} \tag{7}$$

where the lower and upper bounds (L and U) for $\mathbf{x}$ are problem specific (*e.g.* $-\infty$ and 0 for log reflectance). $\lambda_{eq}, \lambda_{gt}, \lambda_{lt}$, and λ_s are weight parameters. Once we solve the quadratic system for the values of $\mathbf{x}$, we floodfill each superpixel with the corresponding value, producing the final output map.

4. Experiments

We use a common framework for both our mid-level tasks: intrinsic image decomposition and depth estimation. The differences in the two models are in the datasets and some parameters, the details of which are given in subsequent sub-sections. The ability to use the same network architecture and optimization model for both datasets is a benefit of our ordinal framework. Furthermore, we can handle images of any input size without rescaling, unlike many other deep network based models.

For both models, we use a patch size of 16×16 pixels around each point. The bounding box around the patches is rescaled to 32×32 pixels. The location masks and ROI are 32×32 pixels in size, and each mask is a Gaussian blob around each of the points, with standard deviation $\sigma = 0.2$

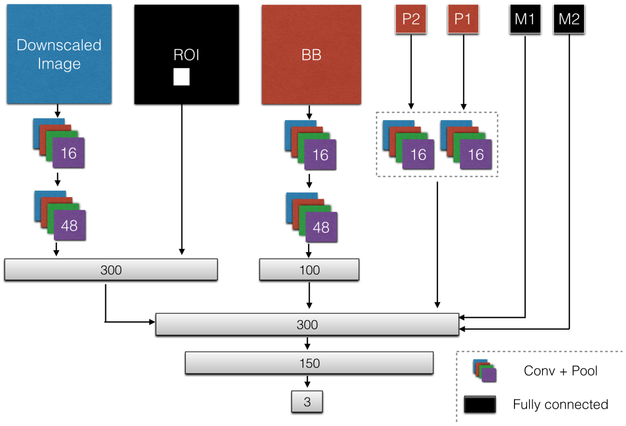

Figure 4: Deep neural network architecture: the numbers in the convolutional layers denote the number of channels. Numbers in fully connected layers denote number of hidden. The convolutional layer's weights after Patch 1 and Patch 2 are shared, forming a small Siamese network. There is a RELU on the output of every layer other than the bottom one, where a SOFTMAX layer is used. All pooling was done on 2×2 neighborhood with a stride of 2. See Figure 3 for details of the input data.

(in normalized 0-1 coordinates). The ROI is a rectangle at the bounding box location within the image, rescaled to the same size of the downscaled image. The image is always downscaled to be 64×64 pixels [1].

We train both the networks using the Caffe framework [16]. The network structure is depicted in Figure 4, having approximately 4 million parameters. Training was performed over $400,000$ iterations with batch size 128. We start with a learning rate of 0.01 and reduce it every $30,000$ iterations (with a reduction rate of 0.1). Momentum was 0.9 and weight decay 5×10^{-4}. We used an NVIDIA GeForce Titan X GPU, on which training takes approximately 10 hours.

4.1. Intrinsic image decomposition

We train our intrinsic image decomposition network on the Intrinsic Images in the Wild (IIW) dataset [5]. Each image in this dataset (5430 images in total) is associated with human annotations for selected pairs of points (i, j). The annotators were asked the question "which of the two points have darker surface reflectance?" and provided one of three responses "i is darker than j" (gt), "i is lighter than j" (lt) or "Equal" (eq). We use these points and annotations as ground truth data for our network. Since the dataset is not split into train and test sets, we generate train, test and validation sets. The test and validation sets have 500 images each, and the rest of the images are put in the training set (4230 images). For each point pair, we extract patches around each point in the pair, location masks, a bounding box covering both patches, a region of interest (ROI) mask within the image and, finally, the downscaled image. Together, this forms the input to our network (see Figure 3).

After training, we solve for reflectance and shading for each image in the IIW test split. We assume grayscale shading and solve for scalar reflectance (as is done in [28, 5]). We use superpixels which are approximately 3% of the image width with long range connectivity (see Figure 2). Weights for the different terms are $\lambda_{eq} = 5$, $\lambda_{gt} = 1$, $\lambda_{lt} = 1$ and $\lambda_s = 0.5$. We set $\sigma_{eq} = 0.001$, $\sigma_{gt} = 0.001$ and $\sigma_{lt} = 0.001$. We set $\mu_{gt} = \log(2)$ and $\mu_{lt} = \log(2)$. The smoothness term is applied to the log shading, that is, we smooth $\log I_i - \log x_i$. This gives a non-zero $\mathbf{b}_s$ smoothness term in Eq. 7. Parameters were adjusted using the validation set. Since we solve for $\log$ reflectance values, the lower bound in Eq. 7 is set to $-\infty$ and the upper bound to 0. After the solution is found we take the exponent of the result to obtain the final reflectance values (between 0 and 1).

For evaluation we follow the procedure introduced in [5] and use the Weighted Human Disagreement Rate (WHDR), the average disagreement rate with human annotators, weighted by their confidence:

$$\mathrm{WHDR}_{\delta}(\ell, \mathbf{x}) = \frac{\sum_{ij} w_{ij} \mathbb{1}(\ell_i \neq \tilde{\ell}_{i,\delta}(\mathbf{x}))}{\sum_{ij} w_{ij}} \tag{8}$$

$$\tilde{\ell}_{ij,\delta}(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{\mathbf{x}_i}{\mathbf{x}_j} > 1 + \delta \\ 2 & \text{if } \frac{\mathbf{x}_j}{\mathbf{x}_i} > 1 + \delta \\ E & \text{else} \end{cases} \tag{9}$$

where $\mathbf{x}$ is the estimated reflectance map, δ is the tolerance level (we use 0.1, as in [5]) and ℓ_{ij} and w_{ij} are the ground truth human annotations and human confidence weights for the ij-th pair. Table 1 shows the performance for our framework, compared to a number of other state-of-the-art methods. We have used the results provided by [5] for comparison to other methods; the parameters for the competing methods were optimized with respect to the *whole* data set including the test images, while ours was fitted only on the train and validation sets). Note that our method significantly outperforms other methods, demonstrating more than a 10% relative decrease in error. Figure 5 shows the recovered reflectance/shading pairs for some images from the test set, along with the results of some of the other methods. Our recovered reflectance maps largely ignore illumination and texture related changes (such as the wooden cabinets and the ceilings).

4.2. Depth from single image

For depth recovery, we train our model on the NYU Depth v2 dataset [25]. Since we cannot directly train on the

[1] All code, datasets and models are at https://dilipkay.wordpress.com/ordinals/

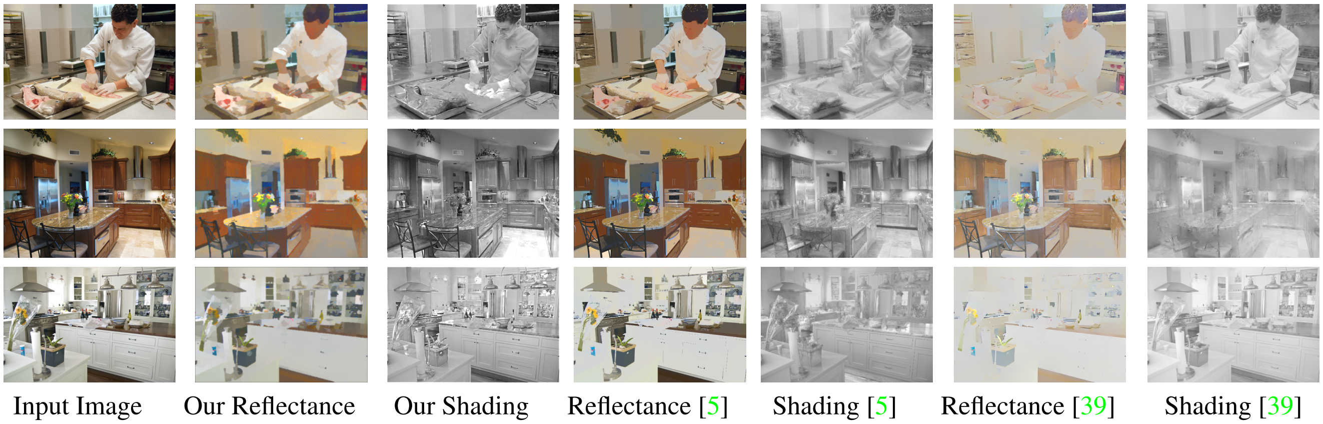

Figure 5: Intrinsic image decomposition results: our recovered reflectance maps largely ignore illumination related changes.

metric estimates provided in the dataset, we have created a set of ordinal relationship annotations for it, analogous to the IIW dataset. For each image in the training set, we extract points and point pairs as described in Section 3. We set the ground truth label ℓ_{ij} for the ij-th pair as follows:

$$\ell_{ij,\delta}(\mathbf{Z}) = \begin{cases} 1 & \text{if } \frac{\mathbf{Z}_i}{\mathbf{Z}_j} > 1+\delta \\ 2 & \text{if } \frac{\mathbf{Z}_j}{\mathbf{Z}_i} > 1+\delta \\ E & \text{otherwise} \end{cases} \quad (10)$$

where $\mathbf{Z}$ is the ground truth depth, and $\delta = 0.02$ is an empirically chosen threshold. After the annotated points and pairs are created, we extract local and global contexts from the RGB images. No metric information is provided during training.

After training, we solve for depth for each image in the NYU test set. We use superpixels which are approximately 3% of the image width with long-range connectivity. Weights for the different terms are $\lambda_{eq} = 1$, $\lambda_{gt} = 1$, $\lambda_{lt} = 1$, and $\lambda_s = 10$. We set the upper and lower bounds in Eq. 7 to 0 and 10. Unlike in the case of reflectance estimation, there is no constraint on the inequality magnitudes for depth differences. Therefore, we set a large variance for $\sigma_{gt} = 4$, $\sigma_{lt} = 4$, and set $\sigma_{eq} = 0.1$. The relatively larger value of λ_s encourages the global solution to be smooth where the image is smooth.

Method	WHDR	WHDR$^=$	WHDR$^{\neq}$
Ours	**17.86%**	**15.99%**	**24.21%**
Dense CRF [5]	20.30%	18.06%	28.66%
Nonlocal [39]	23.26%	17.82%	35.55%
Clustering [11]	25.12%	24.47%	29.09%
RetinexG [12]	26.01%	29.75%	25.99%
RetinexC [12]	26.31%	30.93%	24.47%
Optimization [30]	31.54%	37.29%	24.37%

Table 1: Weighted human disagreement rate (WHDR) for different algorithms. This is the mean disagreement rate of reflectance ratios weighted by human confidence as calculated on the test set. We use $\delta = 0.1$ for the ratio threshold as in [5]. Note that our method outperforms the previous state of the art significantly. The left column is the average over all edges, the middle is over equality edges only and the right column is over inequality edges only. Numbers are somewhat different from [5] because of the train/test split (see text for details).

We compare our metric results on the NYU test set, to that of [17, 3, 8]. The results on the standard benchmarks are reported in Table 2. We also compare our method with outputting just the mean over the training set, and using the mean as the prediction for any test image. The paper of [8] gives details on how the error measures are computed. Due to the specific nature of this dataset, the mean image (a cone of depth, depicted in the supplementary material) has reasonable performance.

Since our network was trained on ordinal measurements, we also introduce an ordinal error measure analogous to WHDR, (Eq. 8), which we dub WKDR (Weighted Kinect Disagreement Rate). WKDR is computed in exactly the same way as WHDR, but with the ground truth and recovered depths maps, from which labels are computed using Eq. 10. We compare our performance under this measure to that of [8], and the results are given in Table 3. Under this measure, our performance gap with [8] is considerably reduced. This suggests that with a different globalization method and more powerful priors, our metric performance on depth estimation could be improved.

In Figure 6, we show some example depth maps from our globalization, on images from the NYU dataset. In the grayscale depth maps, darker intensities are closer (smaller depth value) and lighter intensities are further. We note that we recover the overall subjective depth structure of the scene quite well, often with crisp edges at depth discontinuities.

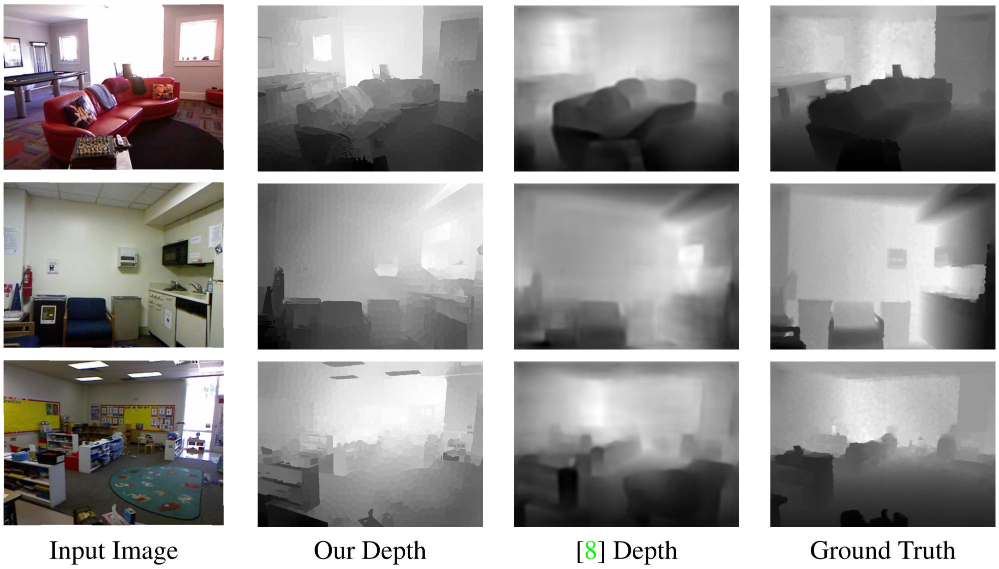

Figure 6: Recovered depth from single images. From left to right: input image, our result, result of [8] and ground truth. While the results of [8] are qualitatively superior, our results do capture major structures of the image, and have crisper edges.

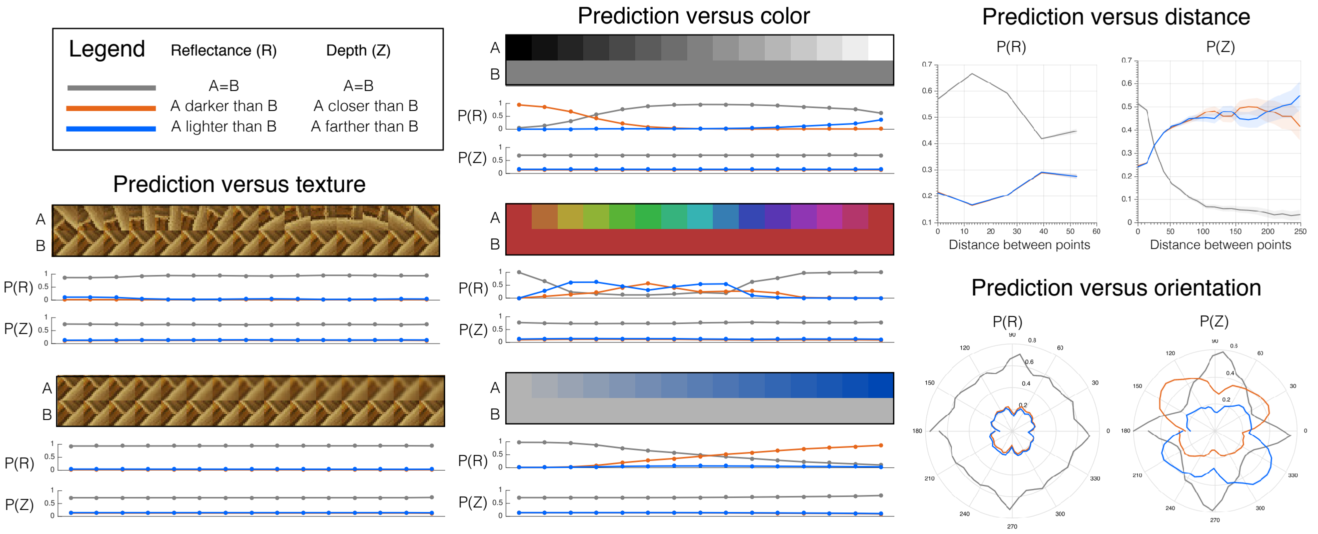

Figure 7: The results of the network analysis experiments: the left two columns measure the dependency of the networks' output to local appearance differences. It can be seen that both networks are invariant to texture deformations (left column) such as blur and rotation. The reflectance network learns that a strong intensity difference means a reflectance change (a "Retinex"-like rule [18], with threshold of about 0.3), though it is sensitive to the polarity of the change. It also predicts that "red" hues are darker than than "green" and that saturated colors are darker than unsaturated, even though there is no intensity difference. Finally, on the right column we see the dependency of the network output as a function of relative distance and orientation of the patches. We see that both the depth and reflectance network have learnt that nearby points tend to be the same reflectance/depth. Additionally, the depth network learns that points closer to the bottom of the image tend to be closer, and that axis aligned pairs tend to be equal. The reflectance network does not show this behavior. See text for further details.

4.3. Network analysis

Deep networks have a large number of parameters, and often, understanding what they learn from data has proven to be challenging. Some recent progress has been made in this aspect [37]. While we can gain partial understanding by looking at learned filters and "network inversions", another approach is to perform controlled experiments on the network outputs.

Our analysis is somewhat analogous to experiments in human psychophysics, where responses to controlled stim-

Method	RMSE	RMSE (log)	RMSE (log s. inv)	absrel	sqrrel
Ours	1.20	0.42	0.47	0.40	0.54
Eigen [8]	**0.75**	**0.26**	**0.27**	**0.21**	**0.19**
Wang [34]	**0.75**	-	-	0.22	-
Liu [22]	0.82	-	-	0.23	-
Li [20]	0.82	-	-	0.23	-
Karsch [17]	1.20	-	-	0.35	-
Baig [3]	1.0	-	-	0.3	-
Mean	1.22	0.43	0.48	0.41	0.57

Table 2: Metric error measures on the NYU test set [25] (lower is better). Some error measures are not available for some of the methods. We refer the reader to [8] for details on the measure computation.

Method	WKDR	$\text{WKDR}^{=}$	$\text{WKDR}^{\neq}$
Ours	43.5%	**44.2%**	41.4%
Eigen [8]	**37.5%**	46.9%	**32.7%**
Mean	47.4%	47.0%	47.6%

Table 3: Ordinal error measures on the NYU dataset (lower is better). On these measures, our method achieves competitive performance with [8] and significantly better than the mean image.

uli are measured. We present a set of experiments which attempt to uncover some of the different rules the networks have learned from both datasets. This is by no means an exhaustive set, but it sheds light on some simple rules. Interestingly, some of the rules correspond to previously proposed algorithms [18, 19].

We measure the response of the network to different appearances of the two input patches. We zero the input of all other channels to measure the effect of the local appearance in isolation. Figure 7 shows the response of the network to differences in intensity, saturation and hue of the patches. The figure shows the different class probability assignment as a function of the parameter changes.

It seems that the reflectance network has learned a "Retinex"-like rule [18] where intensity changes above a certain threshold (0.3 here in [0,1] scale) are considered a reflectance change. We can repeat the experiment on hue and saturation differences: here, it would seem that the reflectance network learned that saturated colors (even though the intensity difference is 0) appear "darker" than unsaturated colors to humans (as provided in the ground truth data) and that some hues appear darker than others. The depth network's response is invariant to these changes, as would also be expected. We also test the dependence of the network output on rotated and blurred copies of the same textures. Both networks are invariant to this kind of transformation.

We can measure the dependence of the network outputs to the relative location and orientation of the points of interest. We marginalize the response of the network on all pairs of points in the tests, and plot the results as a function of relative orientation and distance. Results can be seen in Figure 7. While the reflectance network is largely invariant to orientation, the depth network learns two rules: the bottom point tends to be closer (a prior that has been used in the literature [19]) and points that are axis aligned (vertical or horizontal) tend to be at the same depth (which is a result of the indoor dataset we used, where walls and cabinets are ubiquitous). A final experiment is the dependence on the distance between the points: as expected, both network learn that nearby points tend to be equal, inline with widely employed smoothness priors.

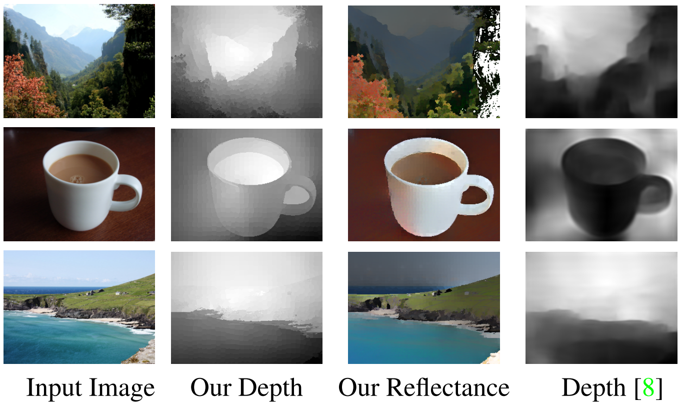

Figure 8: Generalization results on images significantly different than the ones in the dataset. From left to right: input image, our depth map, our reflectance map and depth recovered by [8]. For the depth maps, darker is closer (smaller depth).

4.4. Generalization

To test subjectively the generalization capability of the framework, we used images which are quite different from the training sets, such as close up objects and outdoor scenes. In Figure 8, we show our recovered reflectance and depth maps, and the depth map recovered by [8]. Since ground-truth depth and reflectance maps are not available for these images, no quantitative measure is available.

5. Discussion

We propose a framework that tackles mid-level vision problems by making ordinal decisions about points in an image. We demonstrate that such a system is able to provide good performance while still being interpretable. We are able to utilize the same framework for different modalities such as reflectance and depth. Since the output dimensionality of the classifier is tiny, a small set of "psychophysics" experiments sheds some light on the inner workings of the system. Globalizing these decisions allows to build a full metric explanation of the image to output reflectance, shading and depth maps. Future work will involve allowing different systems to interact to reinforce their decisions.

References

[1] http://www.cnn.com/2015/02/26/us/blue-black-white-gold-dress/. 1

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels. Technical report, 2010. 3

[3] M. H. Baig, V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan. Im2depth: Scalable exemplar based depth transfer. In *WACV*, pages 145–152, 2014. 6, 8

[4] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, pages 17–24, 2013. 1, 2

[5] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM TOG*, 33(4):159, 2014. 2, 5, 6

[6] D. N. Bhat and S. K. Nayar. Ordinal measures for image correspondence. *IEEE PAMI*, 20(4):415–423, 1998. 1, 2

[7] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, 2007. 2

[8] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *arXiv:1411.4734*, 2014. 2, 6, 7, 8

[9] W. D. Ellis. *A source book of Gestalt psychology*, volume 2. Psychology Press, 1999. 1, 2

[10] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *IJCV*, 40(1):25–47, 2000. 2

[11] E. Garces, A. Munoz, J. Lopez-Moreno, and D. Gutierrez. Intrinsic images by clustering. In *CGF*, volume 31, pages 1415–1424, 2012. 6

[12] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342, 2009. 2, 6

[13] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR*, pages 2033–2040, 2011. 1

[14] B. K. Horn. Determining lightness from an image. *CGIP*, 3(4):277–299, 1974. 2

[15] B. K. Horn. Obtaining shape from shading information. In *Shape from shading*, pages 123–171, 1989. 2

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM ICM*, pages 675–678, 2014. 5

[17] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *ECCV*, pages 775–788. Springer, 2012. 6, 8

[18] E. H. Land and J. McCann. Lightness and retinex theory. *JOSA*, 61(1):1–11, 1971. 2, 7, 8

[19] I. Leichter and M. Lindenbaum. Boundary ownership by lifting to 2.1 d. In *ICCV*, pages 9–16, 2009. 8

[20] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, pages 1119–1127, 2015. 1, 2, 8

[21] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *CVPR*, pages 2752–2759, 2014. 2

[22] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. *CVPR*, 2015. 1, 2, 8

[23] D. Marr and A. Vision. A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*, 1982. 1, 2

[24] T. Narihira, M. Maire, and S. X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, pages 2965–2973, 2015. 2

[25] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 5, 8

[26] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE PAMI*, 24(7):971–987, 2002. 1, 2

[27] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510, 2011. 1, 2

[28] C. Rother, M. Kiefel, L. Zhang, B. Schölkopf, and P. V. Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. In *NIPS*, pages 765–773, 2011. 2, 5

[29] J. Sadr, S. Mukherjee, K. Thoresz, and P. Sinha. The fidelity of local ordinal encoding. In *NIPS*, pages 1279–1286, 2001. 2

[30] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR*, pages 697–704, 2011. 6

[31] R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 2

[32] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *IEEE PAMI*, 27(9):1459–1472, 2005. 2

[33] J. T. Todd and J. F. Norman. The visual perception of 3-d shape from multiple cues: Are observers capable of perceiving metric structure? *Perception & Psychophysics*, 65(1):31–47, 2003. 2

[34] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, pages 2800–2809, 2015. 1, 2, 8

[35] Y. Xiong, A. Chakrabarti, R. Basri, S. J. Gortler, D. W. Jacobs, and T. Zickler. From shading to local shape. 2014. 1, 2

[36] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *ECCV*, pages 45–58. Springer, 2012. 3

[37] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014. 7

[38] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE PAMI*, 21(8):690–706, 1999. 2

[39] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE PAMI*, 34(7):1437–1444, 2012. 6

[40] T. Zhou, P. Krahenbuhl, and A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015. 2